International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Ensemble-Based Predictive Framework for Computational Workload Forecasting Using Google Borg Traces

[1]Sanjeev Kumar
Research Scholar, Department of Computer Science and Application, Om Sterling Global University Hisar, India
sanjeev.jangra.in@gmail.com, sanjeevcse221@osgu.ac.in
[2](Dr.) Saurabh Charaya,
Professor, School of Engineering and Technology, Om Sterling Global University Hisar, India
Saurabh.Charaya@gmail.com
[3]Dr. Rachna Mehta
Associate Professor, School of Engineering and Technology, Om Sterling Global University Hisar, India
drrachnamehta@osgu.ac.in

Abstract- This paper introduces a powerful framework based on data regarding future predatory workloads using Google Borg Traces, which is one among the largest examinations of cluster-extensive resource management figures. In the study, the researchers are going to create a predictive model in the form of an accurate and interpretable prediction of CPU utilization using systematic data preparation, advanced preprocessing, feature engineering, and ensemble-based machine learning methods. The data, which was comprehensive in terms of scheduling of tasks, CPU and memory consumption, system performance indicators, were completely cleaned, standardized and converted into formats that are analytical reliable. To select the features, the Mutual Information analysis was utilized, and the Quantile, Power and Z-score transformations were used to normalize the features to improve the stability of the model. Complex feature engineering procedures were used to capture both temporal, frequency and statistical dynamics, and Truncated Singular Value Decomposition (SVD) was used to dimensionality reduce computational efficiency. Three ensemble regression algorithms were used: Random Forest, Light Gradient Boosting Machine (LightGBM), and CatBoost and tested on Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). The best of them was the Random Forest Regressor (MAE = 0.000003, RMSE = 0.000109, R^2 = 0.999915), with LightGBM and CatBoost trailing behind it with very close predictive accuracy. To further improve interpretability, SHAP analysis was used to determine the most significant predictors, including the average and maximum use of CPU, which has a significant effect on the outputs of the model. The findings substantiate that ensemble-based models are effective in modeling non-linear relationships that are very complex, and exist amongst system metrics and resource utilization. Altogether, the work provides a scalable, interpretable, and high-performance model of the intelligent workload prediction and optimization model within large-scale

computing systems that can be useful in enhancing efficiency, reliability, and predictive control in recent cloud and cluster computing systems.

Keywords - Computational Workload Prediction, Google Borg Traces, Ensemble Machine Learning, Resource Utilization Forecasting, Feature Engineering and Explainability.

1. Introduction

E- Governance has become an important tool of enhancing efficiency, transparency, and accessibility of government in the contemporary digital transformation period. [1]. In the world, governments have begun to embrace use of electronic systems to provide necessary services to their citizens in various ways, including online payment of taxes, Internet identity verification, health care management, social welfare transfer, and grievance redress system. Being time sensitive and mission critical, such services need strong infrastructures with the ability to meet huge user demands effectively.[2]. Nevertheless, with the volume and the variety of user interactions ever increasing exponentially, e-Governance systems encounter huge challenges in ensuring smooth performance and avenues of service provision amidst varying workloads.[3]. Unpredictable user traffic, system request and data processing loads tend to cause degradation of services, latency or even system failures. [4]. This creates a strong necessity to have intelligent load prediction and management models capable of providing resilience, scalability and reliability to e-Governance infrastructures. The conventional load balancing and performance optimization techniques of e-Governance architectures are mostly based on the use of static threshold-based mechanism of load balancing or rule-based techniques.[5]. Although these techniques are able to cope with the predictable workloads, in dynamic or uncertain environments where the load changes rapidly like in response to policy announcements, financial deadlines or emergency services they tend to fail.[6]. In addition, the traditional models are not flexible, can not learn and cannot process high dimensional and heterogeneous information in real time. This shortcoming demands a paradigm change to AI-based predictive models, which can actively forecast the future load on a system and optimize resource utilization on the same[7]. Artificial Intelligence (AI) and Machine Learning (ML) may be applied to the prediction of loads, which may have a large impact on the decision-making process, finding more intricate patterns, learning on past cases, and predicting future demand peaks more accurately[8]. The proposed study is concerned with design and development of a new load prediction model that would enhance the resiliency and efficiency of e-Governance systems. The model also uses the state-of-the-art AI methodologies, including deep learning, ensemble models, and time-series predictions, to make precise predictions about the workload of the system based on numerous influencing variables, including, but not limited to, user request trends, service type, time interval, and network performance parameters. Through intelligent predictive means, the system is capable of dynamically establishing computational resources, efficiently scheduling tasks and ensuring optimal performance rates even in the high demand times.[9]. The method is not only effective in ensuring that the system resources are utilized effectively, but also reduces service downtimes, response delays, and costs of the operations. The proposed model placed special focus on the notion of resilience that is the capacity of an e-

Governance service to predict, withstand, and recycle quickly in case of disruptions[10]. The application of proactive load predictions, intelligent response to faults, and adaptive control of the system is what resilience is accomplished in this context. The infusion of AI makes it possible to ensure that the system is constantly tracked in terms of performance indicators, identify deviations and implement corrective measures in response to these indicators.[11]. Such prediction and curing power leads to the e-Governance infrastructure being more resilient to crashes, computer attacks, or sudden user outbursts. This type of framework is essential in sustaining the confidence of the people and uninterrupted access to digital services particularly during emergencies or during operations that are critical. Also, the model analyzes data fusion and hybrid learning algorithms to enhance the predictive accuracy through fusing multiple data sources and analytical algorithms[12]. The application of the hybrid AI models that integrate deep learning and statistical forecasting methods assist in capturing the short-term and long-term load behavior trends. These observations can be used to create scalable and adaptive prediction mechanism applicable in various applications of e-Governance.[13]. The proposed solution is also compatible with cloud-based and edge computing platforms, which means that the model may be implemented in the distributed architecture to facilitate localized decision-making and real-time analytics. The study is expected to develop a holistic AI-powered load prediction framework that will optimize the performance, reliability as well as resilience of e-Governance systems[14]. The proposed model will change the traditional e-Governance infrastructures into smart self-optimizing ecosystems by incorporating predictive intelligence, adaptive management of resources, and resilient architectural design. Effective implementation of this model can become a major move towards the achievement of the concept of smart governance with technology-based decision-making to achieve quicker, more dependable, and citizen-focused service delivery to citizens. In the end, the study will help in establishing sustainable and future-proof e-Governance platforms that can serve the dynamic digital needs of the society.[15].

2. Literature Review

Lilhore 2024 et al. Load balancing is enhanced in cloud data centers by integration methods of virtual machines (VMs). Nevertheless, the balance between cost, performance, quality of service, and adherence to service-level agreements is still difficult to find. Changes in workload and maintaining resources allocated are yet to be managed successfully due to limited resource-level provisioning. This study proposes a dynamic workload allocation model that is hybrid in cloud computing. The model works in two steps. One, optimization techniques are applied to optimize system settings in order to enhance prediction and optimization. Then forecasting framework compares the resource use over time and examines workload statistics to identify patterns. The method can address the issue of the load balancing and over-provisioning by utilizing numerous resources simultaneously. A wide range of experiments with the Google cluster traces indicate that the framework can be effectively used to enhance the allocation of resources and workload management, as well as cloud performance. The proposed solution is hugely precise, accurate and distributes the load

efficiently to ensure that the cloud resources are utilized to the maximum and the needs of the services are also met.[16].

Chandrakanth 2024 et al. With the development of cloud computing, I am interested in the dynamically changing workload computing, storage, and network resources allocation. The paper will focus on forecasting workload, resource utilization, and real-time goals to enhance resource assignment, Quality of Service, and decrease the total costs by a substantial margin. I follow an approach that monitors the performance of a system and makes a guess at what will be received in the future to make intelligent decisions on resource allocation. The system helps to avoid underutilization and over-provisioning through preemptive calculation of resource demands and shifting allocations. The experimental analysis of the true cloud traces shows that our technique is superior to the traditional rule-based and heuristic solutions, resulting in increased resource utilization and a minimization in Quality-of-Service breaches. Through maintaining performance at the optimal and ensuring that it is always up to the expectations of the clients, the dynamic resource allocation strategy makes cloud computing environments more efficient, effective, and competitive[17].

Anuja 2024 et al. E-Governance platforms increase the interaction of government with its citizens, businesses, as also with institutions by making it more transparent, efficient, and accessible. The growing digitalisation of government services creates a vulnerability that requires robust infrastructure that can stay resilient to disruption and recover fast whilst providing uninterrupted service in the face of cyberattack, system outages, or natural disasters. Viable e-governance infrastructure will be based on cloud integration, scalable architecture, disaster recovery system, data encryption and cybersecurity frameworks. The consideration of strategic design and enablers of technology would guarantee continuity of operation and safeguarding of sensitive governmental and citizen data. Inter-departmental cooperation between IT departments, policy developers, and cybersecurity experts enhances online resilience. Lessons and practical practices are learned through the case studies of countries in which resilient systems are implemented. Data protection is guaranteed by regulatory compliance with such standards as GDPR and Digital Personal Data Protection Act in India. Health, education, finance, and identity verification are all critical services including robust infrastructure and therefore technological preparedness and disruption readiness will offer secure, scalable, and reliable e-governance platforms around the globe.[18].

Shaikh 2023 et al. Smart grids involve resource management that needs correct energy prediction to be utilized in demand-side management, optimal dispatch, and load shedding. One aspect that is still a challenge is effective management of the forecasts with a small error of prediction. Time series forecasting with recurrent neural networks is often employed, but issues such as vanishing gradients and low memory recall make recurrent neural networks less useful with sequential data. Temporal convolutional networks offers the solution through parallel processes of outputs, which takes less computation time and the accuracy is also high. Time-dependent energy prediction with temporal convolutional networks shows a better performance when compared to the conventional long short-term memory models, which are

expressed as MAD and sMAPE. Such a strategy will maximize prediction abilities, will guarantee more credible and effective growth of energy sources, be in a better position to face the issue of computational efficiency, and forecasting accuracy and this strategy is highly suitable in dynamic and complex smart grid settings.[19].

Selvan 2023 et al. Workload prediction is a key factor in cloud data centers (CDCs) which guarantees scalability and resource elasticity. The noise, redundancy and poor performance can influence prediction accuracy. The hierarchical tree-based deep convolutional neural network (T-CNN) model with sheep flock optimization (SFO) solves these challenges to increase the power efficiency and workload prediction. A kernel approach is used to preprocess historical data in CDCs and SFO is used to optimize T-CNN weight parameters. The resulting TCNN-SFO methodology minimizes unwarranted power consumption and it predicts the incoming demand with accuracy. The use of the Saskatchewan HTTP traces and NASA benchmark dataset to perform performance evaluation shows that performance evaluation using these traces and data is much better than the existing methods showing improved prediction accuracy and reduced energy consumption on various metrics. The model is found to be efficient as implemented in a Java environment. These findings show that hierarchical T-CNN with SFO is an effective workload forecasting and energy efficient method in CDCs to offer a robust solution to resources management and optimized performance[20].

Table 1 Literature Summary

Authors/year	Methodology	Research gap	Findings
Ahamed/2023 [21]	FEDQWP uses Deep Q-Learning for VM placement optimization.	Existing studies overlook energy efficiency and SLA adherence.	FEDQWP improves CPU utilization, reduces migration, energy, SLA violations.
Lohumi /2023 [22]	Structural analysis of load balancing.	Load unbalancing reduces system performance.	Taxonomy improves cloud resource distribution.
Abdulaziz/2022 [23]	AI-driven IoT techniques applied to analyze and enhance e-Government services.	Existing e-Governance systems face security, adoption, and efficiency challenges.	Proposed framework improves service accessibility, security, and stakeholder satisfaction.
Naqvi/2021 [24]	Analysis of cloud computing, IoT, and 5G impact on digital systems.	Limited understanding of integrated digital systems across all societal levels.	Cloud computing and IoT enable scalable, accessible, versatile digital infrastructures.

Ebrahimzadeh/2020[25]	Empirical study using TAM, questionnaires, SPSS, AMOS, structural equations.	Lack of insight into factors influencing user adoption in banking.	Perceived usefulness and ease shape attitudes; attitudes drive intentions.

3. Research Methodology

The research methodology to be used in this study is a thorough and data-intensive methodology that seeks to predict computation workloads based on the Google Borg Traces data. It is a bulk of data with finer details on the cluster-level resource management and a sound base to be used in the analysis of system performance and predicting the patterns of resource utilization. To make sure that the methodological design is accurate, interpretable, and the model is also computationally efficient, the systematic data preparation, sophisticated preprocessing, feature engineering, dimensionality reduction, and machine learning modeling are used. The data collection and cleaning process started with the extraction of the core data, which included CPU and memory consumption, task priority as well as the scheduling information. Missing, invalid or inconsistent values were dealt with with caution to maintain data integrity. Standardization was done on temporal fields and structured columns were analyzed into numerical sub features so as to be more analytically appropriate. Then the preprocessing of data involved target definition, elimination of multicollinearity and feature selection using the Mutual Information analysis to get the most informative predictors. Quantile, Power and Z-score normalizations (and other feature scaling methods) were used to ensure that the variables have similar distributions. High-level feature engineering Temporal and frequency and rolling statistical features were introduced into the feature engineering to capture the dynamic workload features, and Truncated SVD was used to reduce the dimension to make the computation computationally efficient. LGBMRegressor and SHAP analysis was used to explain the feature importance and model explainability. Lastly, MAE, RMSE, and R 2 became metrics used to implement, fine-tune, and compare ensemble-based regression models, namely, Random Forest, LightGBM, and CatBoost. This staged approach model provided the development of a formidable, interpretable and performing predictive framework of resources loading in large scale computing environments.

3.1 Data Collection

The data that will be used in this research is the Google Borg Traces is complete large scale record of cluster level resource management activities. It covers various system metrics such as job scheduling, job priority, CPU usage, memory usage, and performance metrics. Data was loaded, inspected, and verified in an orderly manner to guarantee consistency, structural integrity, and completeness of available features creating a solid basis on which to subsequently data preprocess and build the model.

3.2 Data Preparation

The complete Google Borg Trace data was carefully filtered to extract a subset of core attributes to concentrate on the variables as they were most relevant to predicting computational workload. These were scheduling information, task priorities, CPU and memory usage, system level performance indicators, including the number of cycles per instruction, and the number of memory accesses per instruction. The invalid or missing values of key fields were either deleted or fixed to provide the reliability of data, and the timestamps were standardized into the consistent forms of datetimes to provide the reliability of time. Nested and array-like columns were read in to separate numerical sub features, to summarize important statistical measures. Categorical variables like cluster identifiers were coded accordingly, redundant variables were removed and the index of the dataset was reinstated to create a clean and structured data that would be smarter in terms of analytical modeling.

3.3 Data Preprocessing

The mean CPU usage was found to be the target variable during the data preprocessing phase, which is the main dependent measure required to be predicted during the modeling process. The remaining variables were taken as independent attributes that would help in comprehending and estimating the workload behavior. All the attributes were categorised according to data types into numerical or categorical to guarantee the correct and systematic preprocessing. This grouping made it possible to apply the appropriate preprocessing methods like scaling of the numerical variables and encoding of the categorical variables to ensure that the data set is consistent and easy to interpret. A multicollinearity reduction process was undertaken in order to achieve an efficient and reliable model. Correlation analysis was done at the pairwise level in order to identify and eliminate highly correlated predictors with a correlation coefficient exceeding 0.9. The removal of redundant variables reduced the chances of overfitting, contributed to a more stable model, and made every feature provide unique and independent information to the process of prediction. These organized steps of preprocessing have created a refined dataset that is best suited to good training and correct prediction of CPU usage behavior at later machine learning phases.

3.4 Feature Scaling and Normalization

Different normalization methods were used to make sure that the features are scaled equally and stabilize the variance. The distributions had been reshaped to normality using Quantile Transformation, skew corrected using Power Transformation (Yeo Johnson) to allow a Gaussian approximation, and normalized with the unit variance and zero-mean using Z-score Standardization. These preprocessing steps increased model robustness, convergence efficiency and consistency in predictive modeling across different numerical ranges.

Table 1 Mutual Information Scores for Key Numerical Performance Features

Feature	Mutual Information Score
memory_accesses_per_instruction	5.275980
cycles_per_instruction	5.274702

maximum_usage_cpus	5.182627
average_usage_cpus	5.180392
tail_cpu_usage_std	4.980238
cpu_usage_min	4.850824
page_cache_memory	4.821488
resource_request_cpus	4.162014
assigned_memory	3.459981
priority	1.669200
scheduling_class	1.120848

3.5 Dataset Splitting

The processed data set was separated into two subsets training (80%), and testing (20%), to allow learning of the models, optimization of the parameters, and the objective assessment of the performance. Random seed values were fixed so that the results of the model could be reproducible and consistent among many experimental runs which increases consistency of the model evaluation process.

3.6 Advanced Feature Engineering

The engineering of advanced features was used to reflect the temporal dependencies and cyclical workload patterns. Autocorrelation characteristics revealed sequential characteristics of CPU and memory activity whereas Fourier Transform revealed dominant frequency characteristics of periodic fluctuations. Localized variations and workload transitions were identified by the use of the Wavelet Transform through Haar functions. Further, rolling statistical characteristics attracted the resource dynamics of the short run. Any gaps created in these transformations were filled backwards to maintain continuity of data, completeness and consistency of data to be further model trained and evaluated.3.7 Dimensionality Reduction

Truncated Singular Value Decomposition (SVD) was applied to reduce high dimensionality and improve the computational efficiency. The data was projected to a lower-dimensional latent space comprising of the first five components with the highest percentage variance. These features derived by SVD were further complemented with some features which were designed by humans to form a complete hybrid feature space to achieve better model performance and accuracy.

3.8 Feature Explainability

An LGBMRegressor was trained to assess the relevance of features and their ability to predict. SHAP (SHapley Additive exPlanations) analysis was used to explain the behavior of the models, and it would give an insight of the impact and direction of influence of each feature. The SHAP summary plots were improved in terms of transparency so that meaningful factors affecting the predictions of CPU usage are identified with understandable explanations of decision patterns in the model.

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

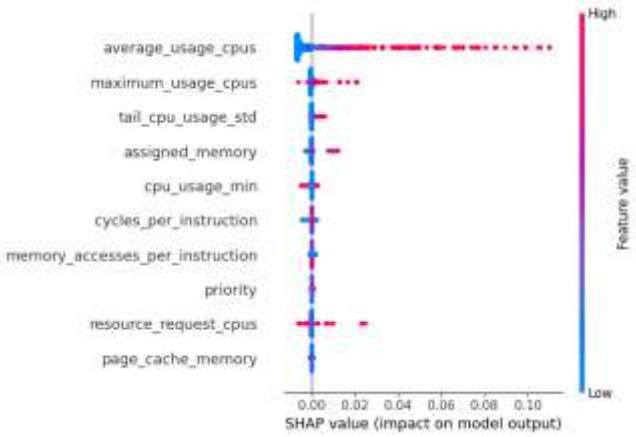

Figure 1 Top features influencing model predictions visualized using SHAP values

This SHAP plot description gives significance to features in a machine learning model. The dots are a SHAP value of a feature on each data point, and indicate the extent to which the feature contributed to the prediction. Attributes such as average usage cpus and maximum usage cpus have the most influence. High feature values are represented as red dots and low as blue dots. The horizontal spread indicates the behavior of the influence and assists to interpret model behavior and select features.

3.9 Dataset Construction Summary

The last training and testing sets were built by combining the best features that were picked using the Mutual Information analysis, engineered time-series and frequency-domain features and the SVD-based dimensionality reduction sub-elements. This universal integration gave a multi-dimensional perspective of workload patterns that integrates statistical, temporal and latent features to improve the predictive quality and interpretability of the machine learning models.

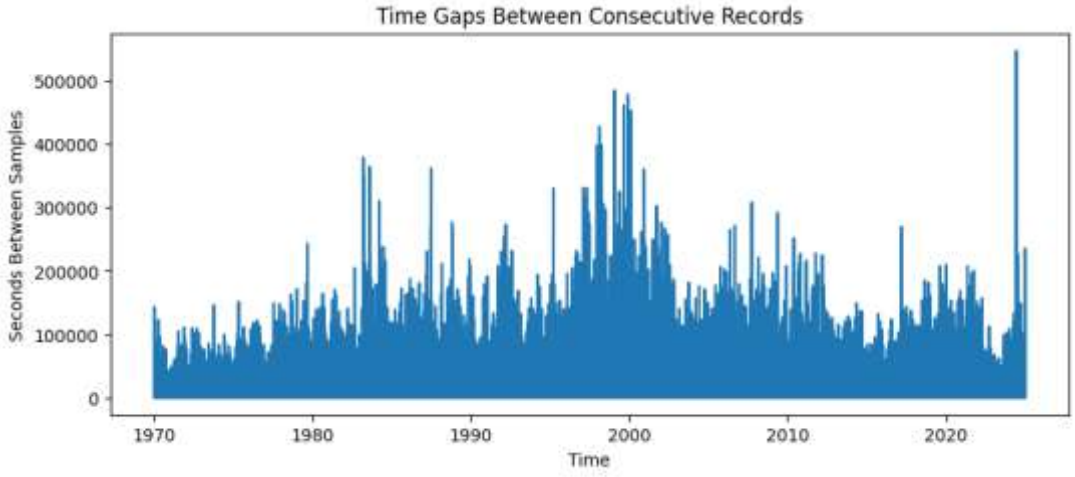

Figure 2 Irregular time gaps highlight changes in long-term data collection patterns

This time bar chart represents the time difference between the sequential records of data in seconds. It covers the period between 1970 and 2020 and indicates the presence of a lot of

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

variability in sampling intervals. The majority of gaps are minor, although there are spikes in the late 1990s, early 2000s, and after 2020, which suggests that there were more gaps in the data collection. Such variations can indicate system downtimes or alteration in the rate of logs, which can provide information on the reliability and consistency of data.

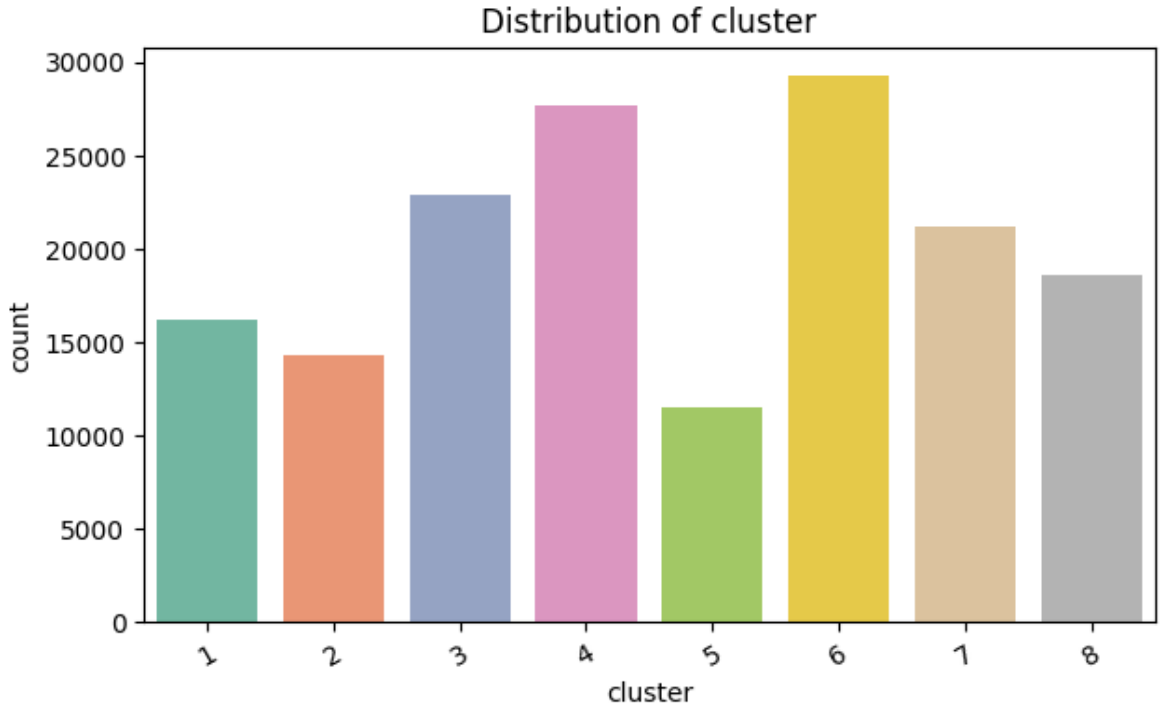

Figure 3 Cluster distribution reveals item count variation across eight distinct groups

The bar chart indicates how the items were distributed in eight clusters. Cluster 6 contains the largest number (about 29, 000), then Cluster 4 and Cluster 3. The lowest number is given in clusters 2 and 5. The bars are color-coded and hence comparing the sizes of clusters is not difficult. This visualization shows imbalances, as well as facilitating the evaluation of the clustering performance in data segmentation.

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

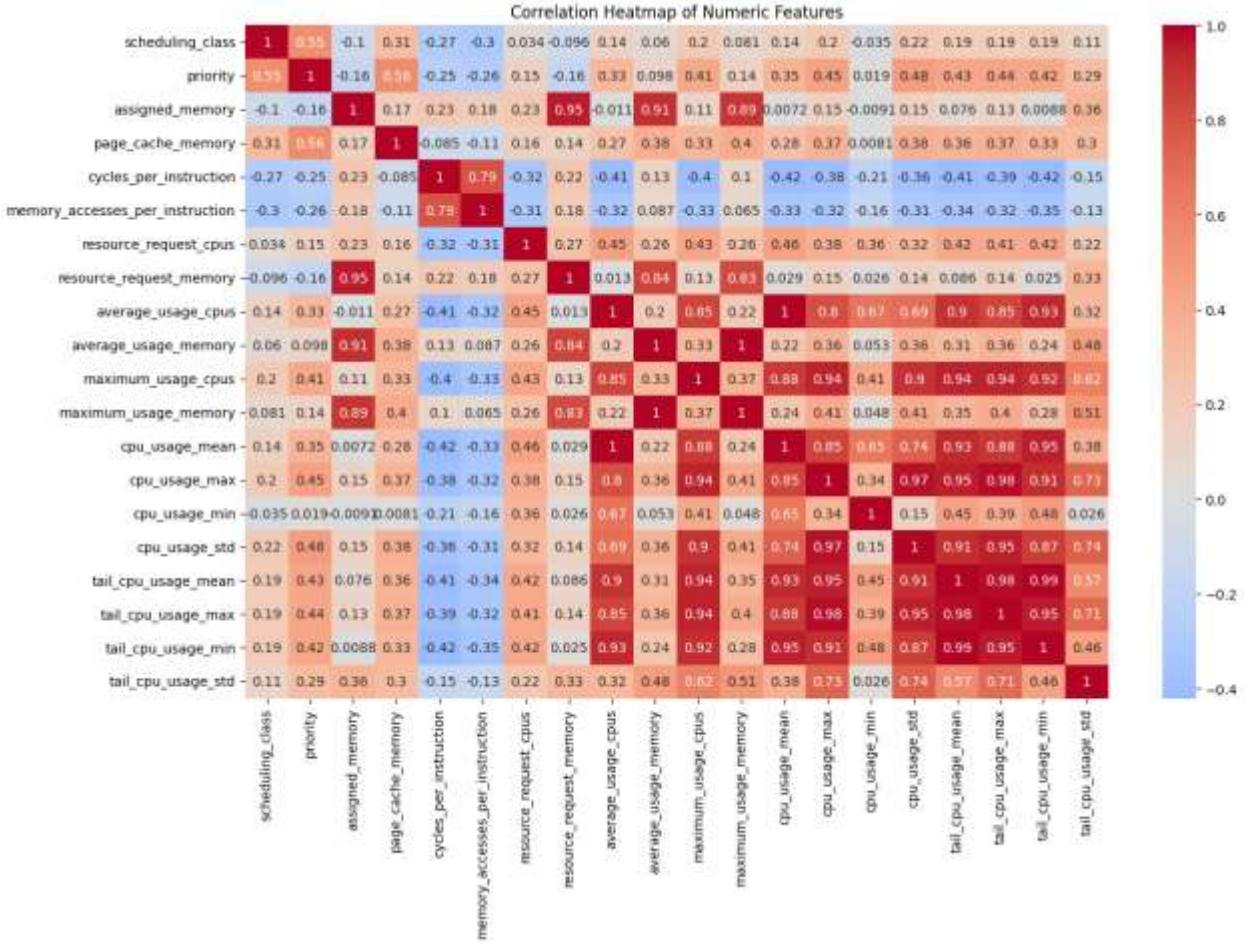

Figure 4 Feature correlations highlight strong CPU usage relationships across multiple metrics

The heatmap is a correlation graph that represents the relationships between different metrics of the system performance. The correlation coefficient between two features is displayed in each cell and is equal to -1 (strong negative), +1 (strong positive). Strong positive correlation is represented by red, negative by blue and weak or no correlation by white. It is interesting to note that variable such as average-usage-cpu, cpu-usage- mean, and maximum-usage-cpu have high positive correlations implying that they move in the same direction. On the other hand, usage metrics are less associated with priority and scheduling class. This heatmap assists in eliminating the redundant features, directing the dimensionality reduction, and enhancing the aspect of model interpretability by showing which variables to be the most associated in the dataset.

International Journal of Research and Technology (IJRT)
International Open-Access, Peer-Reviewed, Refereed, Online Journal
ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529
| An ISO 9001:2015 Certified Journal |

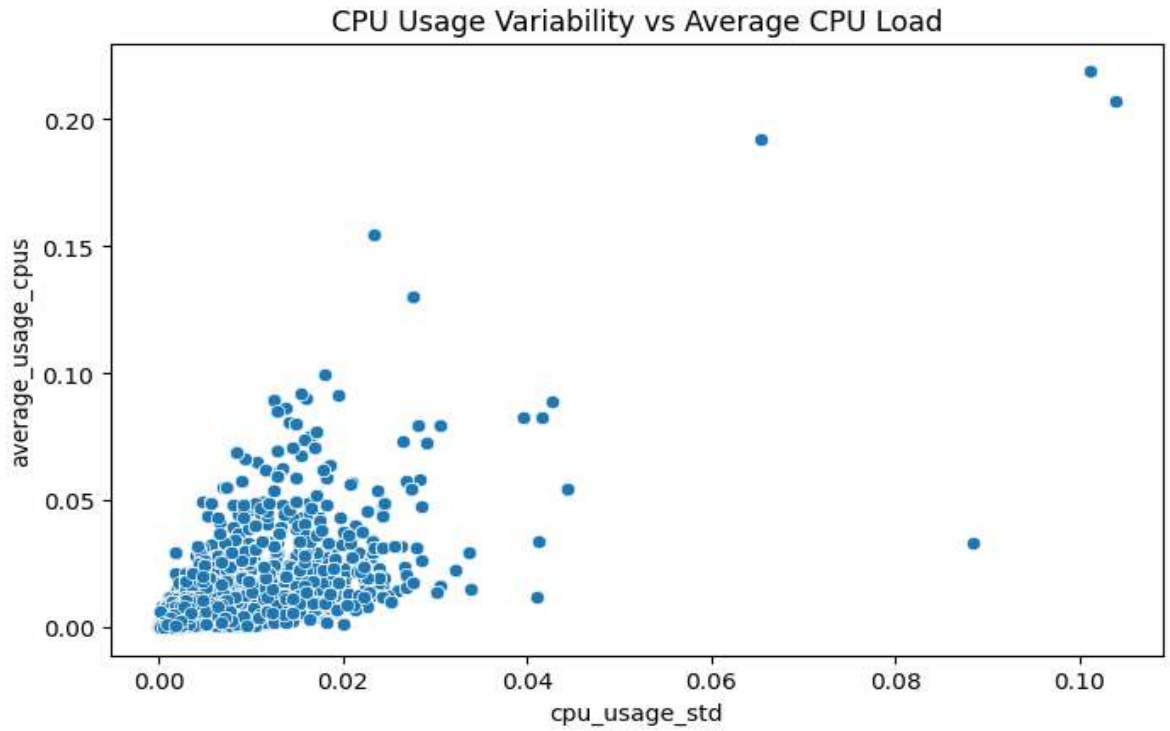

Figure 5 Higher CPU variability often aligns with increased average CPU load

The scatter plot indicates the correlation between variability of the CPU usage and the average CPU load. The blue dots represent the points of data, and the majority of them is located at the lower left meaning low variation of the data and low average use. There are some outliers towards the right which perhaps indicate greater variability and increased CPU load in some cases. The trend is used to determine bottlenecks in performance or inefficiency of resources.

3.10 Machine Learning Modeling Framework

To predict the target variable in this step, the three advanced ensemble based regression algorithms of Random Forest Regressor, LightGBM and CatBoost Regressor were used to ensure accuracy and robustness and flexibility to changing workload patterns. Random Forest Regressor is based on the bagging technique and builds several decision trees with random data sets and combines their results. This combination method reduces overfitting and results in the detection of complex and non-linear associations among resource use measures like CPU and memory use. Light Gradient Boosting Machine (LightGBM) is a lightweight gradient boosting model that uses histogram-based algorithms and leaf-wise tree-building and can therefore train much faster and has lower memory requirements. It is especially appropriate to large datasets such as the Google Borg Traces, due to its scaling capabilities and the need to have performance and computational efficiency. CatBoost Regressor Yandex developed the CatBoost Regressor and it is designed to work with categorical variables intrinsically, and uses an ordered boosting mechanism to prevent target leakage. It effectively parameterically describes complicated feature interactions with little parameter tuning and

high regularization. All in all, these models were optimized and tested against each other to determine the best methodology of predicting dynamically, the multi-resource workloads with reasonable precision and still be interpretable and computationally efficient

3.11 Model Evaluation Protocol

At this stage, a high-precision and robust prediction of the CPU usage was performed by implementing three ensemble-based regression algorithms Random Forest Regressor, LightGBM, and CatBoost. The performance of each of the models has been measured against three major measures: Mean Absolute Error (MAE), Root mean Squared error (RMSE) and Coefficient of Determination (R 2). Random Forest made use of more decision trees to embrace un-linear and complex patterns and minimize overfitting. LightBGM is a fast and scalable gradient boosting framework, which enhanced computational efficiency with the help of histogram-based learning and leaf-wise learning. CatBoost was a strong predictor of categorical variables with ordered boosting that minimizes leakage of targets and maximizes interpretability. These models were biased and compared systematically according to the set performance measures to determine which might be the most efficient and reliable method of predicting multi-resource workloads in large scale computing systems.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

4. Results and Discussion

The outcome of the experiment proves the outstanding predictive capabilities and trustworthiness of the ensemble-based regression models used in this study to predict CPU usage via the use of the Google Borg Traces dataset. Three major key performance measures Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Coefficient of Determination (R 2) were used to evaluate the accuracy, consistency and the explanatory power of the model in order to provide a balance score of the three parameters. The Random Forest Regressor was the most accurate and consistent model, and it had the lowest MAE and RMSE scores; and the highest R 2 measure. This finding indicates high potential of Random Forest to learn non-linear and intricate interrelations between multiple system parameters like CPU utilization, memory consumption and scheduling attributes and reduce overfitting by its ensemble averaging process. Light Gradient Boosting Machine (LightGBM) too was keen to show excellent results with relatively low error rates and high prediction accuracy. Its histogram-based and leaf-wise optimization made it computationally efficient and converged quickly in particular when used in large scale datasets. CatBoost, albeit with lower accuracy,

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

had a consistent good performance since it has the advantage of automatically accommodating categorical variables and strong regularization strategy that contributed to the aspects of better interpretation of the model and minimized the aspect of bias. In general, the predictive accuracy of all three models was very high with their R 2 values of more than 0.999. This implies that the models could explain practically the whole variance of CPU usage depending on the chosen and engineered characteristics. The inclusion of temporal, frequency-domain, and dimensionality-reduced characteristics played a significant part in the strength of these findings. These results confirm the usefulness of the ensemble learning methods as resources utilization forecasting methods and also provide a robust methodological platform to achieve intelligent workload management and predictive optimization systems in a large-scale computing setting.

Table 2 Comparison of Model Performance Metrics for Predictive Accuracy

Model	MAE	RMSE	R^2
Random Forest	0.000003	0.000109	0.999915
LightGBM	0.000039	0.000127	0.999884
CatBoost	0.000135	0.000317	0.999286

The table provides the comparison of three machine learning models Random Forest, LightGBM and CatBoost in terms of MAE, RMSE, and R 2. Random Forest has the lowest error rates (MAE: 0.000003, RMSE: 0.000109), and the highest R2 (0.999915), meaning that the predictive accuracy of the model is great. The next in line is LightGBM and CatBoost has a relatively higher error and relatively a low R 2. In general, each of the models performance is strong, although the most accurate and reliable is the Random Forest.

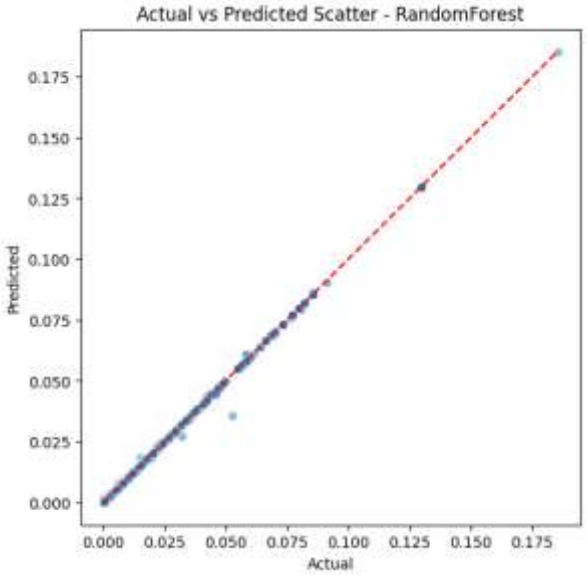

Figure 6 Predicted values closely align with actuals, confirming RandomForest accuracy

International Journal of Research and Technology (IJRT)

International Open-Access, Peer-Reviewed, Refereed, Online Journal

ISSN (Print): 2321-7510 | ISSN (Online): 2321-7529

| An ISO 9001:2015 Certified Journal |

Scatter plot presents the actual and the predicted values of a RandomForest model. The blue dots are the data point, which are very close to the red dashed line on the diagonal (y = x) that means that it is highly predictive. A small variation of the line indicates that the model is able to accurately represent the underlying data trends.

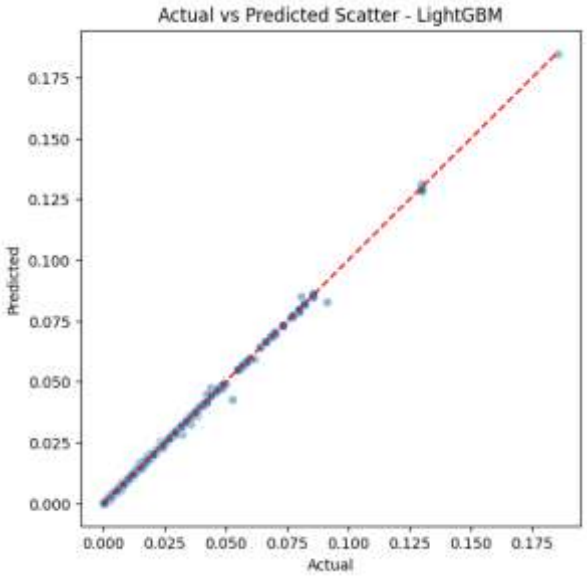

Figure 7 LightGBM predictions closely match actuals, indicating strong model performance

The scatter plot shows the real and the predicted values of the LightGBM model. The blue dots are the data points and the majority of them are near the red dashed diagonal line (y = x) which means that they have a high degree of predictive accuracy. The close-up between the line indicates that the model is very effective and has captured the underlying patterns of the data with the least error.

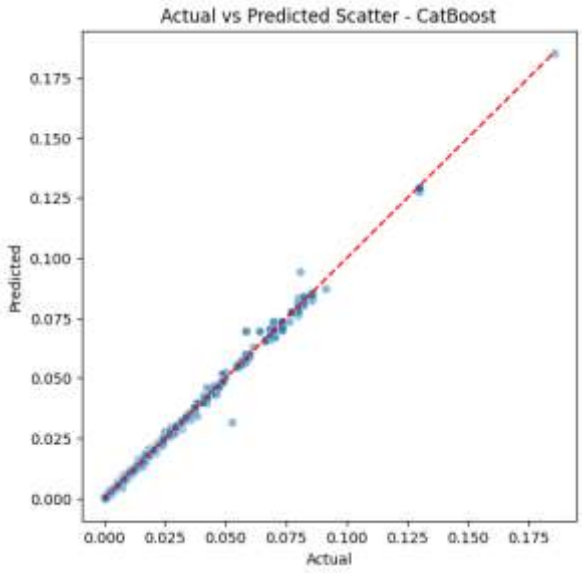

Figure 8 CatBoost predictions align well with actuals, confirming reliable model output

The scatter plot shows observed and forecasted values of the CatBoost model. The blue dots are the points of data and the majority are concentrated around the dashed red line (y = x) which indicates the high level of predictive direction. Although rather scattered as compared to other models, the general tendency is that the CatBoost provides decent predictions with a moderate error.

5. Conclusion

Conclusively, the present paper was able to create an extensive and smart predictive model of computational workload forecasting intensifying with the Google Borg Traces data, showing the strength of the ensemble-based machine learning strategies in processing high-dimensional data of the system volume. To achieve a reliable analytical process and computational efficiency the research was designed based on a methodological framework that involved data cleaning, preprocessing, feature selection, normalization, feature engineering, and dimensionality reduction. Mutual Information analysis was efficient in establishing most informative predictors whereas sophisticated normalization methods like Quantile, Power and Z-score transformations normalized the data to give the best model performance. The temporal and dynamic characteristics of workload behavior were captured in feature engineering and Truncated Singular Value Decomposition (SVD) cut the complexity of the data and conserved the important variance. Random Forest, LightGBM and CatBoost ensemble regression models were applied and compared based on MAE, RMSE and R 2 values to measure their predictive power. Among them, the Random Forest Regressor performed the best (MAE = 0.000003, RMSE = 0.000109, R 2 = 0.999915), then LightGBM and CatBoost in that order, which confirms the strength and scalability of an ensemble approach at workload prediction. SHAP analysis was also beneficial in improving interpretability, and it was observed that average and maximum CPU usage had the highest impact on model predictions which provide clear information on the importance of a feature. These results confirm that it is possible to use ensemble-based regression models to provide an effective way of modeling complex, non-linear relationships between systems metrics and resource utilization. In general, the research offers a high-performance, interpretable, and reliable solution of intelligent workload management in large-scale cloud and cluster computing systems with great implications to predictive optimization, automation of the resource allocation process, and a proficient performance monitoring of the system and, ultimately, leading to the development of the data-driven infrastructure management and the computational intelligence in the current distributed computing systems.

References

[1] S. AlZu'bi, F. Quiam, A. M. Al-Zoubi, and M. Almiani, "Neural Network Architectures for Secure and Sustainable Data Processing in E-Government Systems," *Algorithms*, vol. 18, no. 10, p. 601, 2025, doi: 10.3390/a18100601.

[2] O. E. Aboulqassim, F. Embarak, S. Jayashree, and A. Eltheeb, "Deep Learning-Driven Forecasting Models for Iot Data in Cloud Computing Environments: Leveraging Temporal Convolutional Networks," *J. Theor. Appl. Inf. Technol.*, vol. 103, no. 6, pp. 2108–2122, 2025.

[3] Y. Wang and X. Yang, "Intelligent Resource Allocation Optimization for Cloud Computing via Machine Learning," *Adv. Comput. Signals Syst.*, vol. 9, no. 1, pp. 55–63, 2025, doi: 10.23977/acss.2025.090109.

[4] T. Le Duc, C. Nguyen, and P. O. Östberg, "Workload Prediction for Proactive Resource Allocation in Large-Scale Cloud-Edge Applications," *Electron.*, vol. 14, no. 16, pp. 1–36, 2025, doi: 10.3390/electronics14163333.

[5] A. Rossi, A. Visentin, D. Carraro, S. Prestwich, and K. N. Brown, "Forecasting workload in cloud computing: towards uncertainty-aware predictions and transfer learning," *Cluster Comput.*, vol. 28, no. 4, pp. 1–20, 2025, doi: 10.1007/s10586-024-04933-2.

[6] S. S. Sefati, A. M. Nor, B. Arasteh, R. Craciunescu, and C. R. Comsa, "A Probabilistic Approach to Load Balancing in Multi-Cloud Environments via Machine Learning and Optimization Algorithms," *J. Grid Comput.*, vol. 23, no. 2, 2025, doi: 10.1007/s10723-025-09805-6.

[7] H. Chaudhary, G. Sharma, D. K. Nishad, and S. Khalid, *Advanced queueing and scheduling techniques in cloud computing using AI-based model order reduction*, vol. 28, no. 1. Springer Netherlands, 2025. doi: 10.1007/s10791-025-09581-7.

[8] U. K. Lilhore *et al.*, "Cloud-edge hybrid deep learning framework for scalable IoT resource optimization," *J. Cloud Comput.*, vol. 14, no. 1, 2025, doi: 10.1186/s13677-025-00729-w.

[9] A. B. Kathole *et al.*, "Novel load balancing mechanism for cloud networks using dilated and attention-based federated learning with Coati Optimization," *Sci. Rep.*, vol. 15, no. 1, pp. 1–15, 2025, doi: 10.1038/s41598-025-99559-8.

[10] Raviteja Guntupalli, "Predictive cloud resource management: Developing ml models for accurately predicting workload demands (CPU, memory, network, storage) to enable proactive auto-scaling. AI-driven instance type selection and rightsizing. predicting spot instance interruptio," *World J. Adv. Res. Rev.*, vol. 26, no. 2, pp. 880–885, 2025, doi: 10.30574/wjarr.2025.26.2.1522.

[11] "International Journal of Intelligent Systems - 2025 - Ali - Energy-Efficient Resource Allocation for Urban Traffic Flow.pdf."

[12] Z. Xu, Y. Gong, Y. Zhou, Q. Bao, and W. Qian, "Enhancing Kubernetes automated scheduling with deep learning and reinforcement techniques for large-scale cloud computing optimization," p. 175, 2024, doi: 10.1117/12.3034052.

[13] M. Abouelyazid, "Deep-Hill: An Innovative Cloud Resource Optimization Algorithm by Predicting SaaS Instance Configuration Using Deep Learning," *IEEE Access*, vol. 12, no. July, pp. 92573–92584, 2024, doi: 10.1109/ACCESS.2024.3423339.

[14] K. Jia, J. Xiang, and B. Li, "DuCFF: A Dual-Channel Feature-Fusion Network for Workload Prediction in a Cloud Infrastructure," *Electron.*, vol. 13, no. 18, pp. 1–24, 2024, doi: 10.3390/electronics13183588.

[15] S. Meera and K. Valarmathi, "Optimizing cloud resource allocation: A long short-term memory and DForest-based load balancing approach," *J. Intell. Fuzzy Syst.*, vol. 46,

no. 1, pp. 2311–2330, 2024, doi: 10.3233/JIFS-234054.

[16] S. Simaiya *et al.*, "A hybrid cloud load balancing and host utilization prediction method using deep learning and optimization techniques," *Sci. Rep.*, vol. 14, no. 1, pp. 1–18, 2024, doi: 10.1038/s41598-024-51466-0.

[17] Chandrakanth Lekkala, "AI-Driven Dynamic Resource Allocation in Cloud Computing : Predictive Journal of Artificial Intelligence , Machine Learning and Data Science AI-Driven Dynamic Resource Allocation in Cloud Computing : Predictive Models and Real-Time optimization," *Res. gate*, vol. 2, no. 2, 2024.

[18] A. Chauhan, "Designing Robust and Scalable Infrastructure Solutions to Ensure High Availability and Security in E- Governance Platforms," vol. 2, no. 6, pp. 1–4, 2024.

[19] A. K. Shaikh, A. Nazir, N. Khalique, A. S. Shah, and N. Adhikari, "A new approach to seasonal energy consumption forecasting using temporal convolutional networks," *Results Eng.*, vol. 19, no. March, p. 101296, 2023, doi: 10.1016/j.rineng.2023.101296.

[20] T. Selvan Chenni Chetty *et al.*, "Optimized Hierarchical Tree Deep Convolutional Neural Network of a Tree-Based Workload Prediction Scheme for Enhancing Power Efficiency in Cloud Computing," *Energies*, vol. 16, no. 6, 2023, doi: 10.3390/en16062900.

[21] Z. Ahamed, M. Khemakhem, F. Eassa, F. Alsolami, A. Basuhail, and K. Jambi, "Deep Reinforcement Learning for Workload Prediction in Federated Cloud Environments," *Sensors*, vol. 23, no. 15, 2023, doi: 10.3390/s23156911.

[22] Y. Lohumi, D. Gangodkar, P. Srivastava, M. Z. Khan, A. Alahmadi, and A. H. Alahmadi, "Load Balancing in Cloud Environment: A State-of-the-Art Review," *IEEE Access*, vol. 11, no. November, pp. 134517–134530, 2023, doi: 10.1109/ACCESS.2023.3337146.

[23] A. Al-Besher and K. Kumar, "Use of artificial intelligence to enhance e-government services," *Meas. Sensors*, vol. 24, no. August, p. 100484, 2022, doi: 10.1016/j.measen.2022.100484.

[24] S. A. M. Naqvi, T. Alyas, N. Tabassum, A. Namoun, and H. H. Naqvi, "Post Pandemic World and Challenges for E-Governance Framework," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 10, no. 3, pp. 2630–2636, 2021, doi: 10.30534/ijatcse/2021/1571032021.

[25] F. Mohammad Ebrahimzadeh Sepasgozar, U. Ramzani, S. Ebrahimzadeh, S. Sargolzae, and S. Sepasgozar, "Technology Acceptance in e-Governance: A Case of a Finance Organization," *J. Risk Financ. Manag.*, vol. 13, no. 7, 2020, doi: 10.3390/jrfm13070138.